FRAMEWORK FOR PROCESSING CITIZEN SCIENCE DATA FOR APPLICATIONS TO NASA EARTH SCIENCE MISSIONS

William Teng and Arif Albayrak
ADNET Systems, Inc.
NASA Goddard Earth Sciences Data and Information Services Center
Code 610.2
Greenbelt, MD 20771

ABSTRACT

Citizen science (or crowdsourcing) has drawn much high-level recent and ongoing interest and support. It is poised to be applied, beyond the by-now fairly familiar use of, e.g., Twitter for natural hazards monitoring, to science research, such as augmenting the validation of NASA earth science mission data. This interest and support is seen in the 2014 National Plan for Civil Earth Observations, the 2015 White House forum on citizen science and crowdsourcing, the ongoing Senate Bill 2013 (Crowdsourcing and Citizen Science Act of 2015), the recent (August 2016) Open Geospatial Consortium (OGC) call for public participation in its newly-established Citizen Science Domain Working Group, and NASA's initiation of a new Citizen Science for Earth Systems Program (along with its first citizen science-focused solicitation for proposals).

Over the past several years, we have been exploring the feasibility of extracting from the Twitter data stream useful information for application to NASA precipitation research, with both "passive" and "active" participation by the twitterers. The Twitter database, which recently passed its tenth anniversary, is potentially a rich source of real-time and historical global information for science applications. The time-varying set of "precipitation" tweets can be thought of as an organic network of rain gauges, potentially providing a widespread view of precipitation occurrence. The validation of satellite precipitation estimates is challenging, because many regions lack data or access to data, especially outside of the U.S. and in remote and developing areas. Mining the Twitter stream could augment these validation programs and, potentially, help tune existing algorithms. Our ongoing work, though exploratory, has resulted in key components for processing and managing tweets, including the capabilities to filter the Twitter stream in real time, to extract location information, to filter for exact phrases, and to plot tweet distributions. The key step is to process the "precipitation" tweets to be compatible with satellite-retrieved precipitation data.

These key components for processing and managing "precipitation" tweets (and additional ones to be developed) are not limited to precipitation, nor are they limited to the Twitter social medium. Indeed, to maximize the value of our work for NASA earth science programs, these components should be generalized and be part of an overall framework for processing citizen science data for science research. In this paper, we outline such a framework.

KEYWORDS: citizen science, Twitter, satellite data, earth science, metadata

MOTIVATION AND PREVIOUS WORK

The Twitter social microblogging database, which recently passed its tenth anniversary, is potentially a rich source of real-time and historical, global information for science applications. Over the past several years, as resources permit, we have been investigating the feasibility of extracting from the Twitter data stream--without actively soliciting inputs--useful information for application to NASA precipitation research. There have been similar uses of Twitter (and other social media), mostly related to natural hazards monitoring and management. Participants either knowingly contribute ("active") or not ("passive"). Active examples include Did You Feel It? (earthquake, http://earthquake.usgs.gov/data/dyfi/); Snowtweets (snow, http://snowtweets.uwaterloo.ca/); mPING (rain, http://mping.nssl.noaa.gov/); WOW (rain, http://www.bom.gov.au/wow-support/); GEO-Wiki (land cover, http://www.geo-wiki.org/). Passive

examples include those for earthquake (Sakaki et al., 2012; Earle et al., 2011; Crooks et al., 2013); air pollution (Shu, 2014); and rain (Butgereit, 2014; Lwin et al., 2015).

Both passive and active engagements by participating citizen scientists hold potential. For our previous work, in the passive case, we have experimented with listening to the Twitter stream in real-time for "precipitation" and related tweets, applying basic filters for exact phrases (in different languages), extracting location information, and mapping their resulting distribution (Fig. 1). In the active case, we

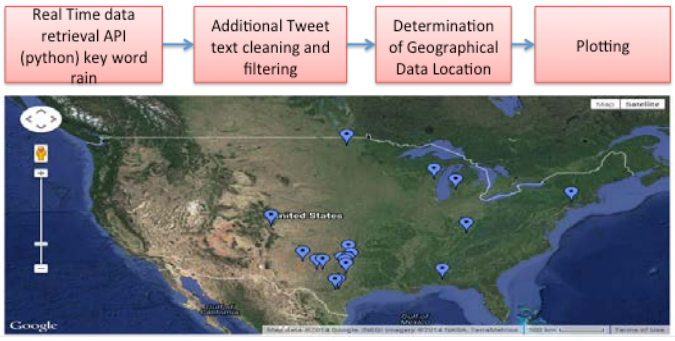

Figure 1. (Top) Basic architecture for processing Twitter data from our previous work. (Bottom) Map of tweets that was part of a middle school science fair project for which we had advised. Our scripts were used by students to handle geolocation of events (e.g., rain, earthquake).

have conducted preliminary experiments to evaluate different methods of engaging with potential participants. For example, we have replied to "precipitation" tweets with Global Precipitation Measurement (GPM) images generated by NASA Giovanni (Acker and Leptoukh, 2007), centered on the tweet locations (Fig. 2). This complementary combination of passive and active engagements is similar to, e.g., the use of both passive and active sensors in the Soil Moisture Active Passive mission (SMAP; https://www.nasa.gov/smap).

Figure 2. Experimental Twitter account (@nasadisc) created to store our replies to "precipitation" tweets. Each reply contained an URL to a NASA Giovanni image centered on the "precipitation" tweet location.

Our general crowd-sourcing strategy is to *not* require participants to explicitly "sign up" or install some app to contribute as citizen scientists, even in the active case. We believe this is a more robust approach. To effectively crowd-source, a large source of crowd, obviously, is needed. The Twitter data stream is such a source of crowd. In a random sampling from our previous work, the number of "precipitation" tweets was about 2,000/min. For especially notable events, the number of tweets per minute can peak orders of

magnitude higher (Krikorian, 2013). In addition, the Twitter "follow" feature, with which followers can see the tweets posted by those they follow, provides a build-in recursive mechanism for recruiting potential new citizen scientists.

POTENTIAL SCIENCE APPLICATIONS

The validation of satellite precipitation estimates across the full range of weather regimes is challenging, because many regions lack data, particularly remote and developing areas. Even where precipitation data are collected, access to the data is often difficult for those developing and validating the underlying algorithms. To improve this situation, the GPM mission, in particular, has pursued dedicated field campaigns and exchanges with a range of international partners (Hou et al., 2014).

Mining the Twitter stream could augment GPM's validation program. The science rationale for mining "precipitation" and related tweets is to potentially obtain a widespread view of rain occurrence. The time-varying set of "precipitation" tweets can be thought of as an organic network of rain gauges. Given tweets of reasonable quality, the two main science issues are (1) the results are mostly only qualitative (no precipitation rate information) and (2) the results are generally only positive (no direct no-precipitation reports), though there are possible ways to mitigate these issues. Currently, the precipitation/no-precipitation boundary is subject to considerable uncertainty, when gauge stations or satellite overpasses are sparse. "Precipitation tweets" could help constrain this boundary. The potential exists for tweets to help tune existing GPM algorithms. Figure 3 shows a rain event in the Los Angeles, CA area on October 23, 2016 and the corresponding distributions of tweets and GPM satellite data.

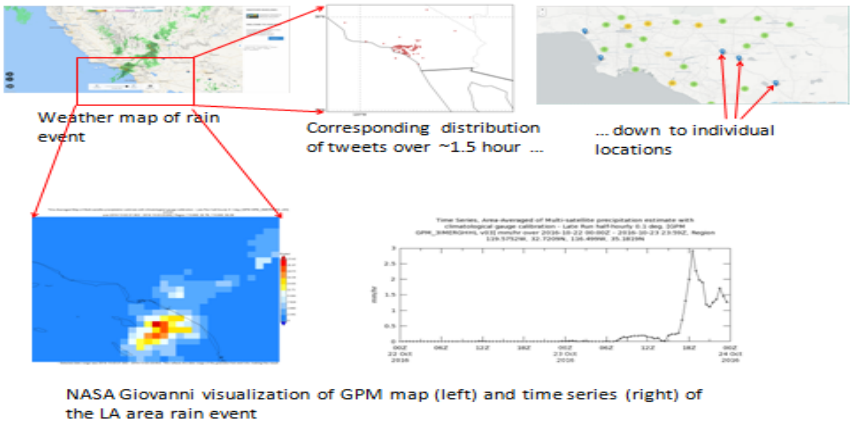

Figure 3. (Top) Weather map of a rain event in the Los Angeles, CA area on October 23, 2016 and corresponding distribution of tweets; (bottom) corresponding GPM precipitation map and time series.

INFRASTRUCTURE FOR TWEET PROCESSING AND ANALYSIS

A key aspect of our Twitter stream analysis infrastructure is its modular design, which provides a flexible algorithm and component plug-in/out capability. This allows for easy updates of algorithms. Likewise, new components can be easily added to the infrastructure. Figure 4 shows our Twitter stream analysis flow, with ongoing work in blue text and planned future work in red text. Notably, though our current work is focused on precipitation and the GPM mission, the infrastructure we are developing will be easily reusable for other phenomena (e.g., air pollution, landslides) and related satellite missions. Furthermore, the overall framework is not Twitter-specific and can be adapted for other social media.

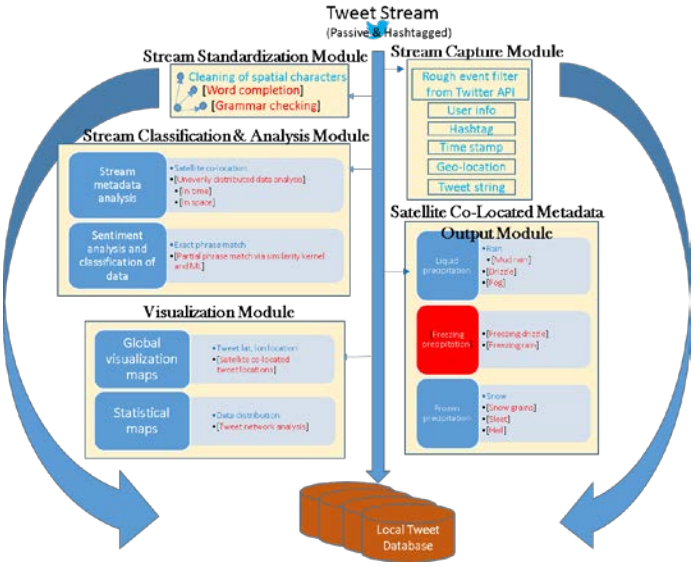

Figure 4. Tweet stream analysis flow (ongoing work in blue text; planned future work in red text).

Four key questions guide our development of tweet processing and analysis infrastructure:

1. How to classify tweets by existence, type, and intensity?
2. How to determine the quality of tweets (e.g., geolocation, reliability of source)?
3. What metadata are needed to enable interoperability with other data sets, to facilitate their use?
4. How to co-locate tweets with satellite data?

Tweet Classification

The purpose of classifying tweets is to retrieve relevant information that could augment existing satellite data validation programs (GPM in our project). This retrieval of information from raw tweets is analogous to the retrieval of geophysical measurements in satellite missions (See Fig. 7). Because the information content of tweets varies in detail, we classify tweets first by existence (i.e., precipitation/no precipitation) and second by type and intensity (e.g., rain, snow, drizzle). Algorithms used for classification also vary, depending on purpose. Currently, we are using phrase matching techniques. Planned for the near future are extensions to similarity measures and learning algorithms (e.g., Support Vector Machines; Campbell and Ying, 2011) (Fig. 5).

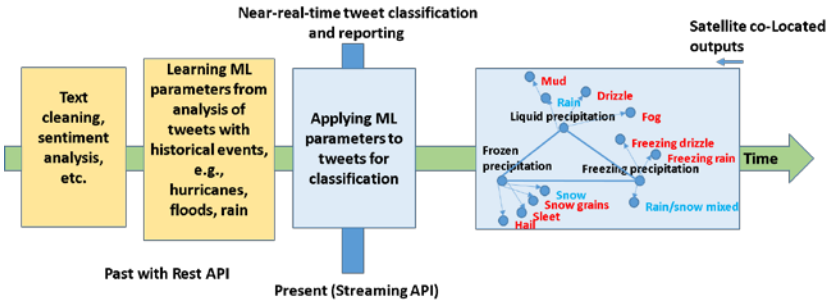

Figure 5. General flow of tweet classification, using machine learning algorithms.

Tweet Quality

The quality of Twitter-derived data is, to a large extent, inherent in the process of tweet processing and applying the data to GPM validation. In terms of processing the tweets, there are multiple factors that could affect the quality of tweets. These include reliability of twitterer (e.g., active, passive, author ranking with

experience level), type of geo-location information (e.g., point, polygon, retrieved from the 140-character tweet), number of retweets, attached images, and links that increase the information level. The quality of the final processed tweets will depend on some kind of weighted combination of these causative factors.

Tweet Metadata

Metadata support the data standardization necessary for interoperability among data systems. Metadata are important for interoperability among relatively similar satellite data systems; but perhaps more so for interoperability among disparate citizen science and satellite data systems. The overall challenge in managing Twitter-derived data is their unconventional nature, relative to data sets in the vast majority of NASA data centers. Specifically regarding the co-location of tweets with satellite data (see next section), certain metadata are needed for valid co-located-with-satellite tweets (e.g., co-located satellite platform, co-location radius, tweet location). Figure 6 summarizes the three types of metadata that will be included in the processed tweets for GPM validation: (1) Raw twitter metadata, (2) retrieved information metadata, and (3) co-located satellite metadata.

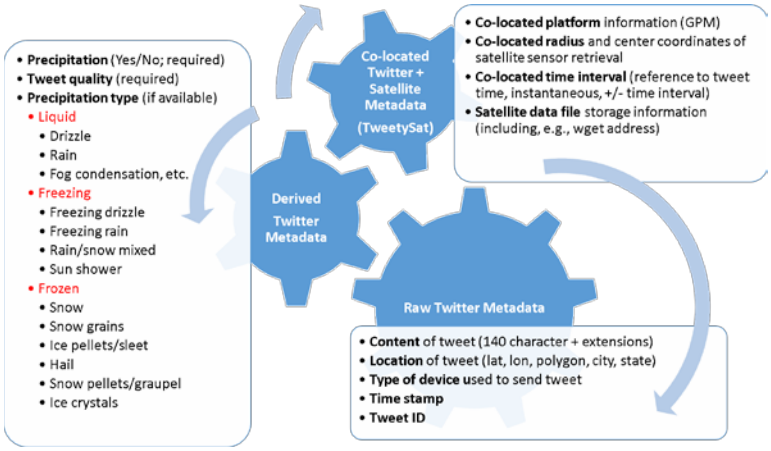

Figure 6. Three types of metadata that will be included in the processed tweets for GPM validation: (1) Raw twitter metadata, (2) retrieved information metadata, and (3) co-located satellite metadata.

Co-locating Tweets with Satellite Data

Before tweets can be inter-compared and analyzed with corresponding satellite data, they need to be accompanied by relevant metadata, co-located with satellite data, and in a compatible format. Figure 7 shows a proposed mapping of tweet processing levels to those defined for NASA satellite data.

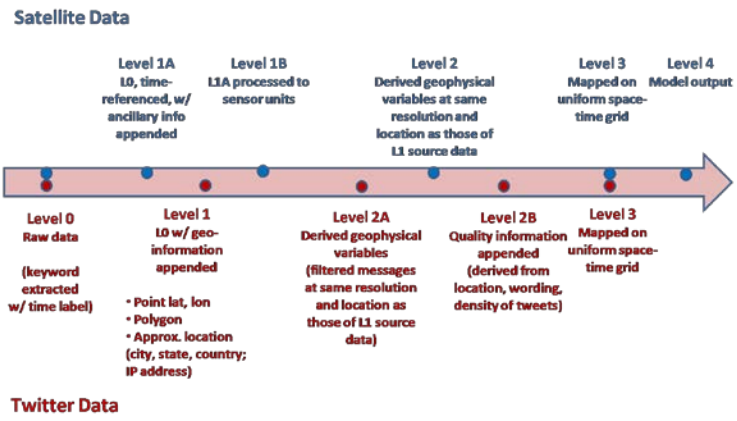

Figure 7. Proposed mapping of tweet processing levels to those defined for NASA satellite data.

The co-location of tweets with satellite data is somewhat more challenging than the more familiar co-location of ground station and satellite data. However, there is an existing Python toolbox, **Op**en **S**ourc**E** **E**OSDIS Level 1,2 **Dat**a Reader Library (OpSEEDat; Albayrak et al., 2014), that collects data from satellites into a multi-sensor data sandbox. We are leveraging OpSEEDat to facilitate the co-location of tweets with the GPM satellite. Figure 8 shows examples of GPM ground track (left), Level 2 sensor data traces (swaths) (bottom), and Level 3 gridded GPM IMERG data (right) (plotted by NASA Giovanni). Figure 9 shows example raw tweets collected during the February 18th rain storms on the U.S. west coast and elsewhere.

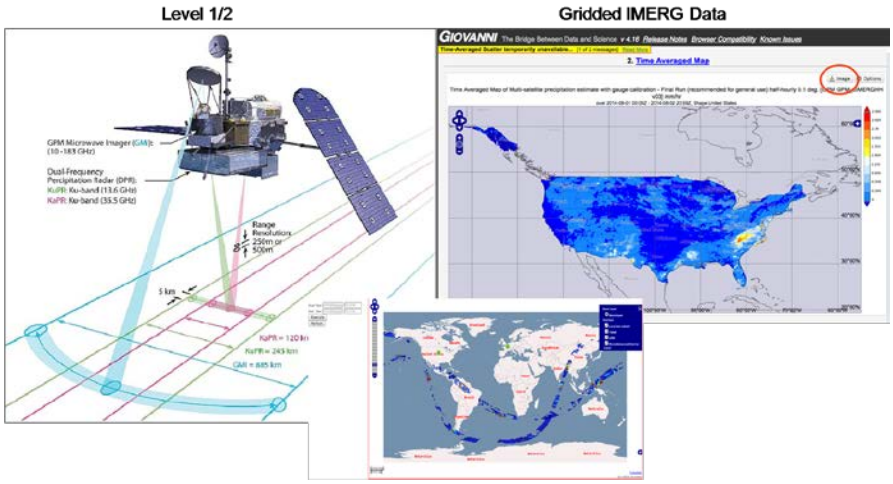

Figure 8. GPM ground track (left), Level 2 sensor data traces (swaths) (bottom), and Level 3 gridded GPM IMERG data (right) (plotted by NASA Giovanni).

Figure 9. Raw tweets collected during the February 18th rain storms on the U.S. west coast and elsewhere. Over a 1.5-hour period, approximately 50,000 precipitation tweets were collected globally (only English), of which 1,500 had exact geo-locations and 2,500 had less exact place information.

Once the GPM satellite data have been processed by OpSEEDat, and the raw tweets have been processed into the same level as that of the satellite data (Fig. 7), an additional step is needed, before OpSEEDat can be used to co-locate tweets with GPM data. The tweets are binned to the same spatial grid as that of the Level 3 IMERG data (Fig. 10). Then, for each grid cell, all the included tweets ("children") are aggregated (in some way) to a representative "parent" tweet (Fig. 11).

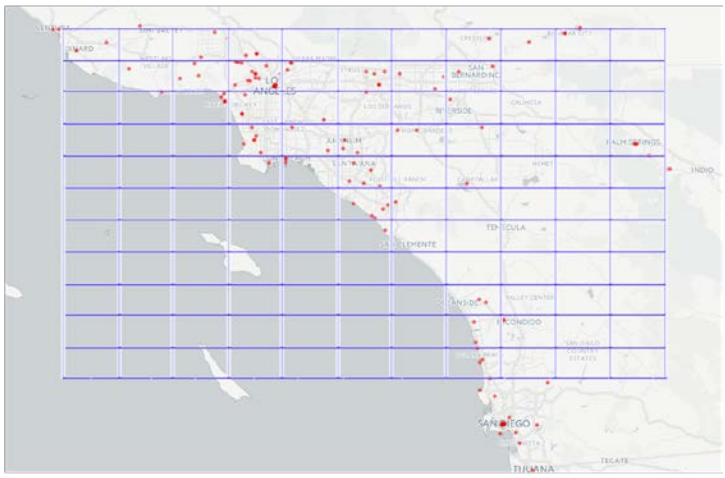

Figure 10. In order to use OpSEEDat to co-locate tweets with GPM data, the tweets are binned to the same spatial grid as that of the Level 3 IMERG data.

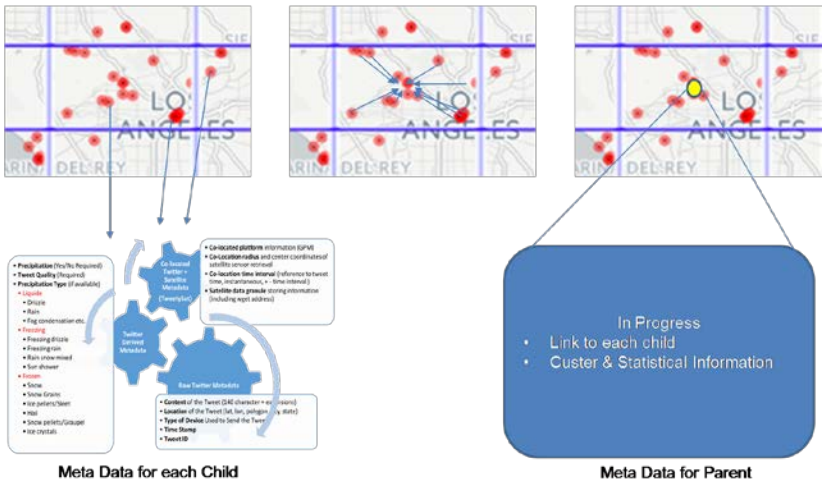

Figure 11. For each grid cell in Figure 10, all the included tweets ("children") are aggregated (in some way) to a representative "parent" tweet. Each "child" tweet (red dot) has its associated metadata (left bottom). Each "parent" tweet (yellow dot) has its associated aggregated metadata (ongoing work).

SUMMARY

Our work to date has demonstrated the potential for extracting from the Twitter data stream useful information for earth science applications, specifically those of NASA earth science missions. Realizing this potential would significantly extend the application realm of Twitter, as a platform for citizen science, beyond natural hazards monitoring. Our recently begun NASA Citizen Science for Earth Systems Program (CSESP) project will enable us to more fully develop our prototype infrastructure for processing tweets

The GPM mission is the focus of this project. But, the project's Twitter stream processing infrastructure is not GPM-specific and is applicable to other satellite missions as well, e.g., Soil Moisture Active Passive (SMAP). Such a demonstration of the value of Twitter to the science research community would bring a new kind of visibility to Twitter and engender a new kind of appreciation of the value of Twitter by these communities.

ACKNOWLEDGMENT

The work reported in this paper was in part supported by a NASA ROSES NNH16ZDA001N-CSESP grant.

REFERENCES

Acker, J.G. and G. Leptoukh, 2007. Online analysis enhances use of NASA earth science data, *AGU EOS Trans.*, *88*(2), 14, 17.

Albayrak, A., B. Vollmer, and A. Savtchenko, 2014. Optimum swath reader library for science data of earth observing system missions (OpSEEDat), NASA Technology Report: GSC-17192-1.

Butgereit, L., 2014. Crowdsourced weather reports: An implementation of the μ model for spotting weather information in Twitter, In: *IST-Africa Conference Proceedings,* pp. 1-9, doi:10.1109/ISTAFRICA.2014.6880593.

Campbell, C. and Y. Ying, 2011. *Learning with Support Vector Machines*, Morgan & Claypool Publishers, San Rafael, 95 pp., doi:10.2200/S00324ED1V01Y201102AIM010.

Crooks, A., A. Croitoru, A. Stefanidis, and J. Radzikowski, 2013. #Earthquake: Twitter as a distributed sensor system, *Trans. GIS, 17*(1), 124-147, doi:10.1111/j.1467-9671.2012.01359.x.

Earle, P.S., D.C. Bowden, and M. Guy, 2011. Twitter earthquake detection: earthquake monitoring in a social world, *Annals of Geophysics, 54*(6), 708-715, doi:10.4401/ag-5364.

Hou, A.Y., R.K. Kakar, S. Neeck, A.A. Azarbarzin, C.D. Kummerow, M. Kojima, R. Oki, K. Nakamura, and T. Iguchi, 2014. The Global Precipitation Measurement Mission, *Bulletin of the American Meteorological Society, 95*, 701-722, doi:10.1175/BAMS-D-13-00164.1.

Krikorian, R., 2013. New tweets per second record, and how! (https://blog.twitter.com/2013/new-tweets-per-second-record-and-how), Twitter Official Blog, August 16, 2013.

Lwin, K.K., K. Zettsu, and K. Sugiura, 2015. Geovisualization and correlation analysis between geotagged Twitter and JMA rainfall data: Case of heavy rain disaster in Hiroshima, *2015 2nd IEEE International Conference on Spatial Data Mining and Geographical Knowledge Services (ICSDM)*, Fuzhou, pp. 71-76, doi:10.1109/ICSDM.2015.7298028.

Sakaki, T., M. Okazaki, and Y. Matsuo, 2012. Tweet analysis for real-time event detection and earthquake reporting system development, *IEEE Transactions on Knowledge and Data Engineering, 25*(4), 919-931, 10.1109/TKDE.2012.29.

Shu, A., 2014. *Data mining of Chinese social media*, PhD. thesis, Rice Univ., Houston, 127 pp.